Minority Language Engineering

Professor Tony McEnery,

Dept. Linguistics and Modern English Language, Lancaster University

Email: a.mcenery@lancaster.ac.uk

Itroduction

- The TEI and corpus building at Lancaster
- MILLE & BIMLLER
- EMILLE -its outline
- Progress on EMILLE
- TEI, NIMLS & BIMLS
- Conclusion

The TEI and Corpus Building at Lancaster

- The use of the TEI on past corpus building projects has shown the scheme to be:
  - Comprehensive
  - Flexible
  - Well suited to linguistic annotation
- In using the TEI we have been able to approach a data of many types.

- Hand-written: The Lancaster/Leverhulme Corpus of Children's Writing
- Transcription of hand-written material
- TEI of use
    - In normalising spelling;
    - In annotating features lost in the transcription;
    - Adding visual annotation;
    - Articulating a multimodal corpus.

See Smith, McEnery & Ivanic, Literary & Linguistic Computing, 1998 (4).

- Speech: The encoding of speech and thought presentation in spoken language
- Transcription and annotation of oral history archives
- TEI of use
    - In encoding linguistic annotations;
    - In helping to track changes and evolving analyses through responsibility statements;
    - Preparing the corpus for presentation as a time-aligned multimodal corpus.

- Historical: The creation of machine readable versions of Early Modern English newsbooks
- Transcription of newsbooks from the Civil War/Commonwealth/Restoration period
- TEI of use

    In normalising spelling;

    In tracking editorial decisions;

    In tracking text reuse across a number of newsbooks.

Minority Language Engineering

- The focus - non-indigenous UK minority languages (NIMLS - McEnery) and British indigenous minority languages (BIMLS - Wilson). Part of Lancaster's focus on widening the range of corpus data available (see McEnery & Ostler, 2000).
- NIMLS - mainly Indic languages and varieties of Chinese, but covering languages such as Arabic and Somali also
- BIMLS - Varieties of Gaelic (Cornish, Erse, Manx, Scots Gaelic, Welsh). We are not covering BIMLS based on English such as Scots and Ullans.

The MILLE Project

- MILLE (Minority Language Engineering)
- Partners: Lancaster University, Oxford University Computing Service
- Steering Group: (Universities) Edinburgh, Sussex and UMIST. (Industry) Canon, Linguacubun, Routledge and Sharp. (Public sector) BBC, ELRA, Dept. Health.
- Funded by the EPSRC (1998 - 1999)
- Pilot project examining the feasibility of constructing NIML corpora

Why?

- Most UK domestic translation tasks are focused on NIMLS and BIMLS
- We are liasing with nations where these are indigenous/major languages
- Yet even where such nations do produce resources, they may not be relevant to the UK context

BIMLLER

- Starting February 2002
- Repeating the MILLE exercise for BIMLs
- Some issues will be similar (code switching), some different (reviving languages, language endangerment), some irrelevant (character encoding).
- Considering the role of such data in preserving dying languages the use of TEI is crucial. We must get the markup right.

Enabling Minority Language Engineering (EMILLE)

- 40 month project funded by the UK EPSRC (grant no. GR/N 19106). Began September 2000.
- Main partners: Lancaster University (McEnery) and Sheffield University (Gaizauskas).
- Others helping (e.g. Oxford)
- Languages initially covered: Bengali, Gujarati, Hindi, Panjabi, Urdu (200,000 word parallel, 500,000 word spoken and 9,000,000 word written corpora each) plus Singhalese and Tamil (9,000,000 word written corpora each)

- Aims:

1.) To generate corpus data for Indic languages

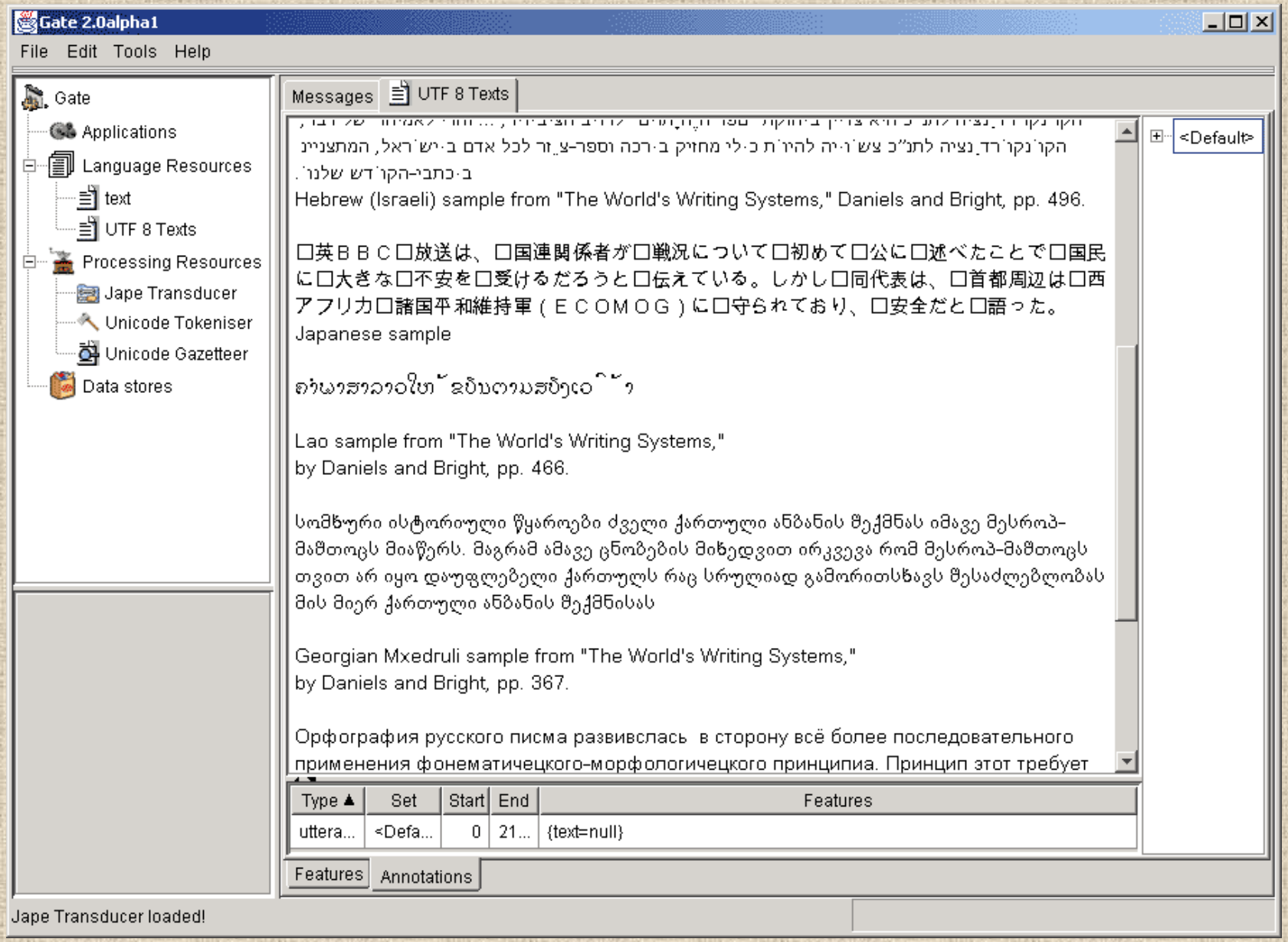
2.) To adapt an existing language engineering architecture (GATE) for NIMLs

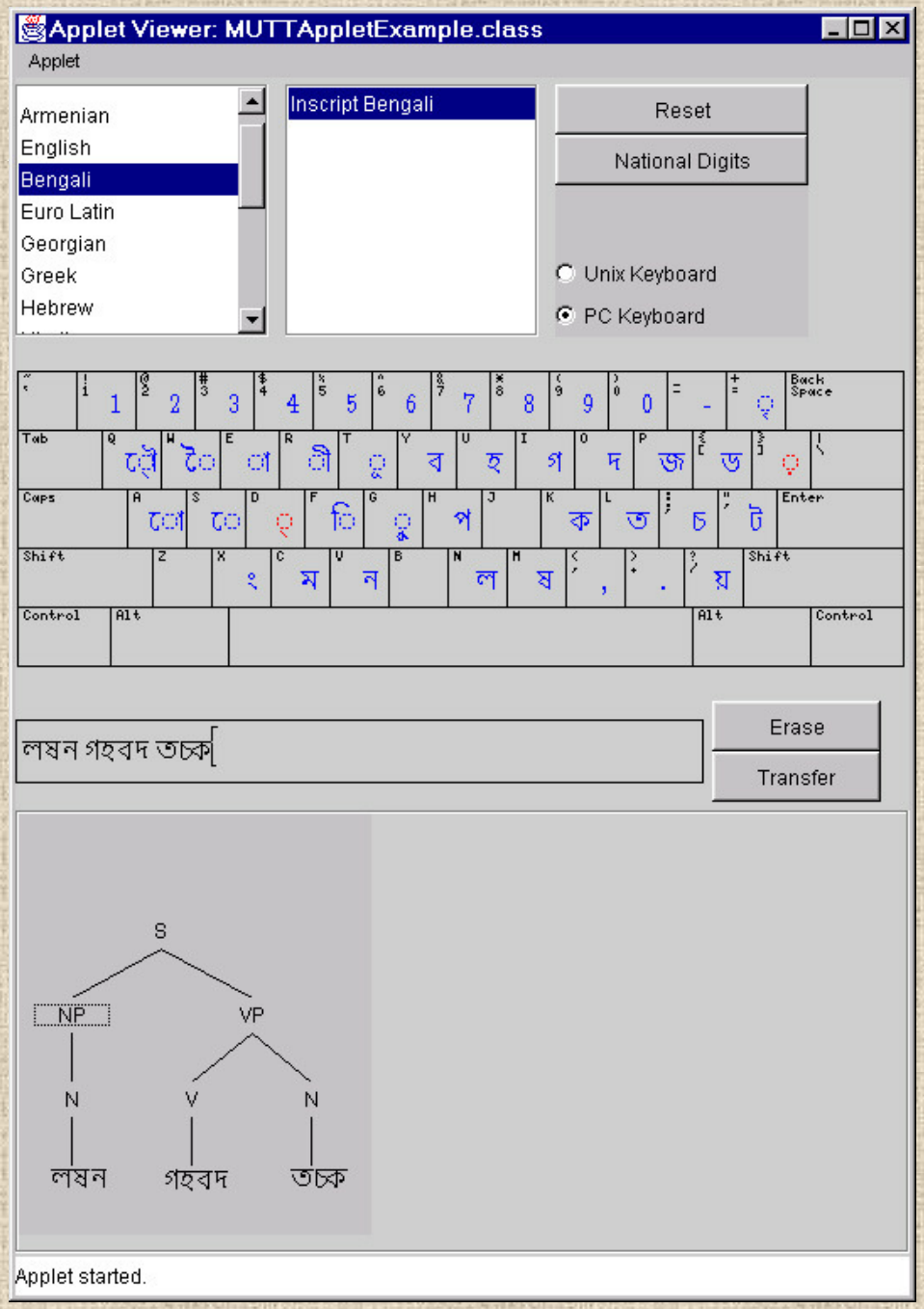
Progress report 1 - data

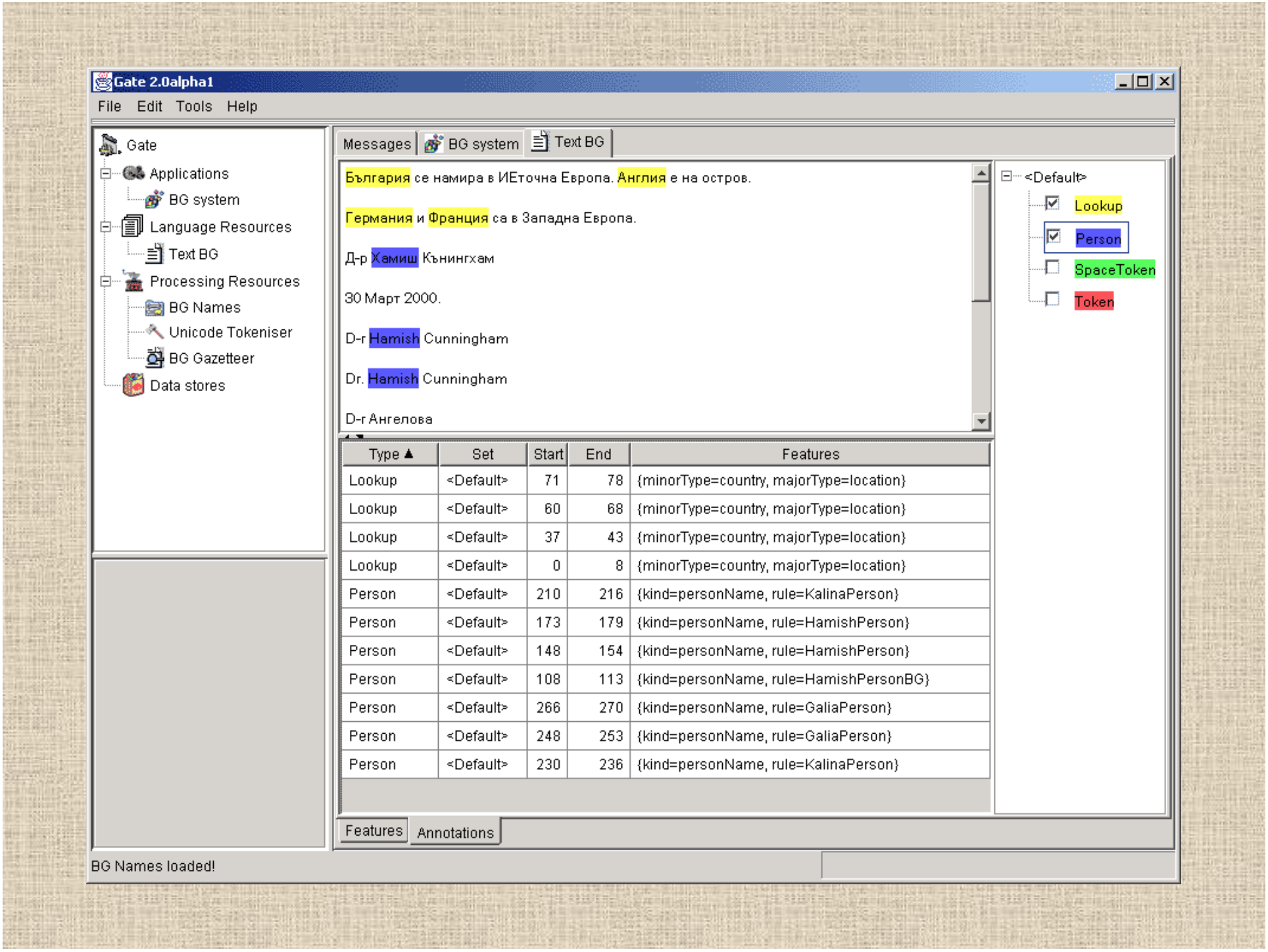
- 24,000,000 words of written data collected to date. We are focusing on news material.
- Collection and orthographic transcription of spoken material on-going. Radio broadcasts main source of data. Around 1,000,000 words transcribed to date. All TEI compliant.
- Parallel corpus material being collected (50,000 words of multiple translations to date)
- Agreement with Central Institute of Indian Languages, Mysore

Progress report 2 - GATE

- Alignment software being embedded within GATE. Part-of-speech tagging for Urdu under development.
- Becoming Unicode compliant in a new Java based version of GATE. Using JMUT from NMCL (cross platform delivery).


Gate 2.0alpha1
File Edit Tools Help
Gate
Applications
Language Resources
text
UTF 8 Texts
Processing Resources
Jape Transducer
Unicode Tokeniser
Unicode Gazetteer
Data stores
Messages
UTF 8 Texts
<Default>
Hebrew (Israeli) sample from "The World's Writing Systems," Daniels and Bright, pp. 496.
□英ＢＢＣ□放送は、□国連関係者が□戦況について□初めて□公に□述べたことで□国民に□大きな□不安を□受けるだろうと□伝えている。しかし□同代表は、□首都周辺は□西アフリカ□諸国平和維持軍（ＥＣＯＭＯＧ）に□守られており、□安全だと□語った。
Japanese sample
Lao sample from "The World's Writing Systems,"
by Daniels and Bright, pp. 466.
Georgian Mxedruli sample from "The World's Writing Systems,"
by Daniels and Bright, pp. 367.
Орфография русского писма развивслась в сторону всё более последовательного применения фонематичецкого-морфологичецкого принципиа. Принцип этот требует
Type
Set
Start
End
Features
uttera...
<Defa...
0
21...
{text=null}
Features
Annotations
Jape Transducer loaded!


Applet Viewer: MUTTAppletExample.class
Applet
Armenian
English
Bengali
Euro Latin
Georgian
Greek
Hebrew
Inscript Bengali
Reset
National Digits
Unix Keyboard
PC Keyboard
Erase
Transfer
S
NP
VP
N
V
N
Applet started.


Gate 2.0alpha1
File Edit Tools Help
Gate
Applications
BG system
Language Resources
Text BG
Processing Resources
BG Names
Unicode Tokeniser
BG Gazetteer
Data stores
Messages
BG system
Text BG
България се намира в ИЕточна Европа. Англия е на остров.
Германия и Франция са в Западна Европа.
Д-р Хамиш Кънингхам
30 Март 2000.
D-r Hamish Cunningham
Dr. Hamish Cunningham
D-r Ангелова
<Default>
Lookup
Person
SpaceToken
Token
Type ▲ | Set | Start | End | Features
Lookup | <Default> | 71 | 78 | {minorType=country, majorType=location}
Lookup | <Default> | 60 | 68 | {minorType=country, majorType=location}
Lookup | <Default> | 37 | 43 | {minorType=country, majorType=location}
Lookup | <Default> | 0 | 8 | {minorType=country, majorType=location}
Person | <Default> | 210 | 216 | {kind=personName, rule=KalinaPerson}
Person | <Default> | 173 | 179 | {kind=personName, rule=HamishPerson}
Person | <Default> | 148 | 154 | {kind=personName, rule=HamishPerson}
Person | <Default> | 108 | 113 | {kind=personName, rule=HamishPersonBG}
Person | <Default> | 266 | 270 | {kind=personName, rule=GaliaPerson}
Person | <Default> | 248 | 253 | {kind=personName, rule=GaliaPerson}
Person | <Default> | 230 | 236 | {kind=personName, rule=KalinaPerson}
Features
Annotations
BG Names loaded!

Progress report 3 - the need for UNICODE

- The main issues we have encountered have related to character interchange
- The writing systems used by Indic languages can be represented in an 8 bit format, but lack of appropriate word processing software has led to a number of conflicting font led solutions to using English-language software, so a may map to A with one font, while mapping to m may map to A with another

Unicode

- The obvious standard - though harmonising to one 8/16 bit representation per writing system is a possibility
- For languages with an 8 bit standard which is widely adhered to this may not seem so necessary
- But for a wide range of languages where 8 bit standardization has not been established/successful it is much more useful

What happens when standardisation fails?

- South Asian languages are good examples of the failure of standardization
- There ARE standards- they are simply not adhered to
- The standards came too late, and now compete with well established rival commercial/shareware standards
- These standards and rivals are mutually incompatible to varying degrees

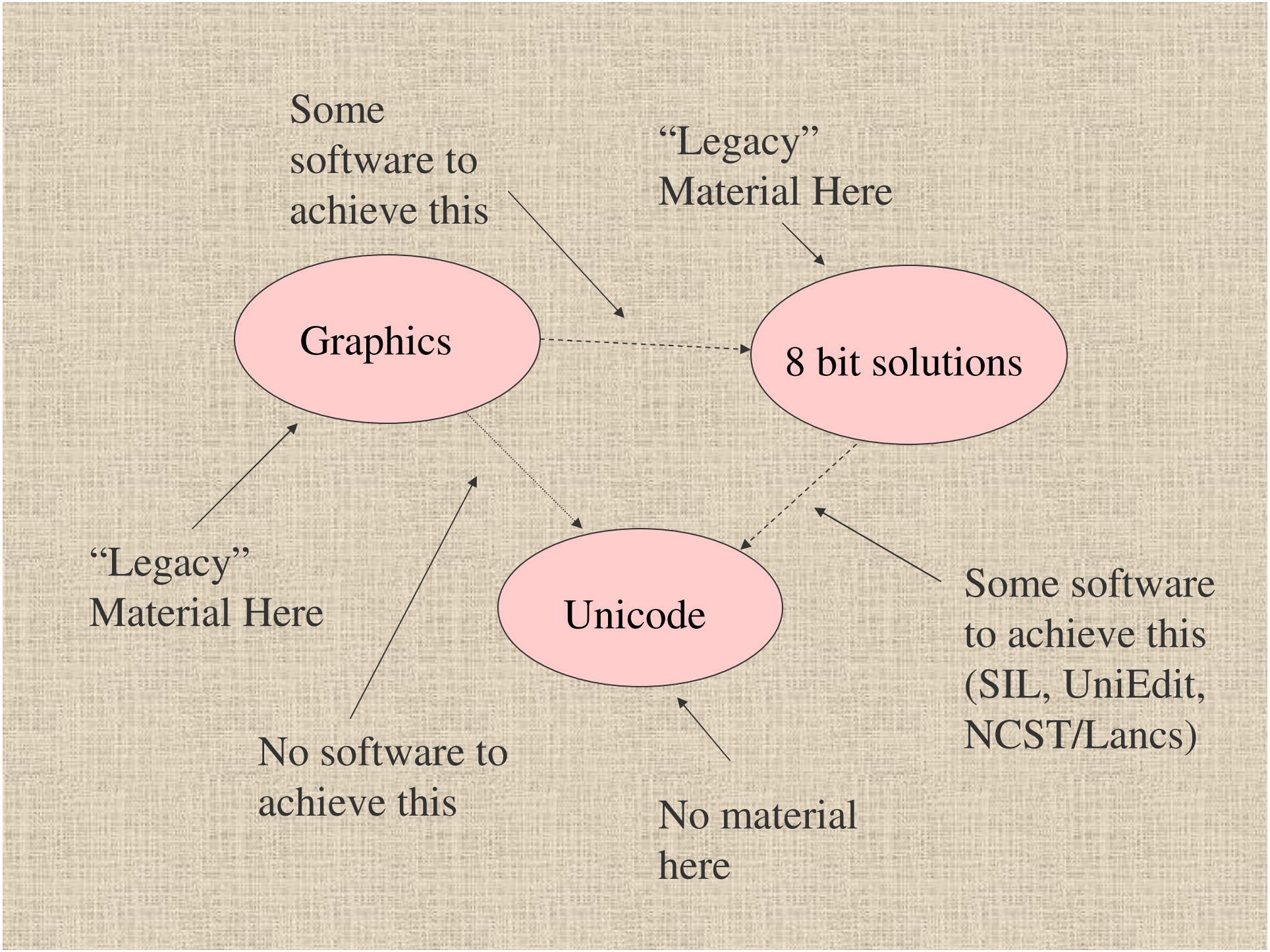
For example, Panjabi (k, g, t)

kgt (Anandpur Sahib, Maboli Systems Inc.)

kgt (Gurbani, Gurbani Foundation)

kgt (Panjabi, Hardip Singh Pannu)

kgt (WCGurumukhi, Duke University)


Some software to achieve this
"Legacy" Material Here
Graphics
8 bit solutions
"Legacy" Material Here
Unicode
Some software to achieve this (SIL, UniEdit, NCST/Lancs)
No software to achieve this
No material here

Solutions?

- TEI WSDs?

```
<character class=lexical>
    <form string='k' ucs-4='0A15'>
      <desc>Gurmukhi letter letter KA</desc>
    </form>
  </character>
```

- UTR 22
- Simple minded ‘bespoke’ programs
- LDC developing ‘best practice’ guidelines in this area

TEI, NIMLS and BIMLS

- Application of the TEI to NIML/BIML data fairly straightforward (Singh, McEnery, and Baker, 2000, 'Building a parallel corpus of English/Panjabi' in Véronis, J. (ed.), Parallel Text Processing : Alignment and Use of Translation Corpora, Kluwer)
- The degree of code switching in some spoken material has led us to use the *distinct* element to allow us to mark this up.

- The degree of borrowing noted may simply be of whole words or whole words with distinct pronunciations (*sap* -> *shop*). However, morphology may be mixed below the word level:
- daktor-e (object)
- daktor-o (locative)
- Using distinct we have worked on a simple scheme to mark up both distinctive pronunciation and morphology in code switching (see Baker, Lie, McEnery & Sebba, 2000).
- Ongoing effort to engage South Asian corpus linguists with the TEI (Burnard, McEnery)

Conclusion

- Use of TEI on-going - indeed just beginning for some languages.
- Work of utility beyond the UK - how well are the NIMLS and IMLS of Europe provided with LE resources? How will TEI be able to help?
- TEI standards applied to data being produced in a wide range of corpus building projects at Lancaster.